Release Guide

Release PR

It is recommended to create a dedicated branch and PR for a new release:

1. Update version in `pyproject.toml`.
2. Run `uv lock --upgrade` to update the lock file.
3. Pull new images on the cluster if inference engine version is updated, more details in next section.
4. Update model tracking information and configs accordingly, more details below.

Update images

If the inference engine version is updated, the Docker GitHub action will build new images tagged by the inference engine version and pushed to DockerHub:

1. Go to `vec-inf-maintenance/maintenance`.
2. Run `pull_image.sh $ENGINE_NAME $VERSION`, e.g. `pull_image.sh vllm 0.14.0`. This will pull the new image to `/model-weights/vec-inf-shared`, and update the symlink accordingly.
3. Go to `vec-inf-maintenance/deployment`.
4. Run model launch and inference tests to ensure the new images are working properly.

Update model tracking information and configs

For any new release, always sync the latest cached configs (`models.yaml` and `environment.yaml`) on Killarney cluster (`/model-weights/vec-inf-shared`) to the repository:

- If new models were downloaded and/or supported on the cluster, they should be reflected in `MODEL_TRACKING.md`.
- Whenever changes are made to the cached configs on any cluster, the cached configs should be pushed to `vec-inf-maintenance/vec-inf-shared/$CLUSTER_NAME`, e.g. `vec-inf-maintenance/vec-inf-shared/bon-echo`

Create a new release on GitHub

1. Create a new tag and release on GitHub
2. Add description for the new release
3. Documentation for the new version will be automatically deployed

---

Revision #1
Created 2026-02-04 05:08:56 UTC
Updated 2026-02-04 05:34:46 UTC